Mock Exam

Social Media and Web Analytics

Spring 2025

Instructions to Candidates

The exam is graded out of **100 points**. Points for each of the three parts are clearly labelled at the beginning of each. The point allocation for each question is also clearly designated.

The **exam duration is 120 mins**.

Write your Student ID and name on every page of the exam.

Answer each question in the text box directly underneath the question. There are additional blank pages at the end of the exam if you need to cross out your work and start again. Clearly label your answers on any of the extra pages.

Answers to the questions should be **based on the course material presented in lectures, Lab Sections and readings.**

Remember, **your goal is to communicate**. Full credit will be given only to the correct solution which is described clearly. **Convoluted and obtuse descriptions might receive low marks, even when they are correct**. Also, aim for concise solutions, as it will save you time and also help you conceptualise the key idea of the problem.

To pass Social Media and Web Analytics, you must receive a grade higher than 5.5/10 for this exam.

Part A: Multiple Choice (30 points)

Answer ***all*** questions. **Each question in Part A is worth 3 points**. There is **no** guessing correction applied to the grading.

Question 1.
Which statement best describes a key difference between the BING and VADER sentiment lexicons?

A. BING provides sentiment scores for individual words, while VADER provides a compound sentiment score for entire texts.

B. BING is optimized for analyzing social media text, while VADER is designed for general-purpose sentiment analysis.

C. BING does not account for sentiment intensity, while VADER accounts for intensity using heuristics and punctuation.

D. BING incorporates contextual understanding of phrases, while VADER relies solely on individual word sentiment scores.

Question 2.
A company implements an A/B test with treatment assignment at the user level and intends to analyse data at the purchase level so that there might be multiple purchases per individual during the experiment. When analysing the data via linear regression analysis, what assumption should they make on the variance of the error term?

A. It is homoskedastic

B. It is clustered at the geographic level

C. It is clustered at the user level

D. It is heteroskedastic

Question 3.
Consider the following regression equation:

$$y_{it} = \beta_0 + \beta_1 Treatment_i + \beta_2 After_t + \beta_3 Treatment_i \times After_t + \varepsilon_{it}$$

Assuming the parallel trends assumption holds, the Difference in Differences estimate of the average treatment effect is

A. $\beta_1 + \beta_3$

B. β_3

C. $\beta_3 - \beta_2$

D. $\beta_3 - \beta_1$

Question 4.
Fill in the blanks. When testing motivations for posting in social media, researchers define Intrinsic Utility as _____(i)____ and predict that experimentally increasing the number of followers would ____(ii)____ if this was the main driver of posting behaviour.

A. (i) inherent satisfaction of posting; (ii) increase the number of user posts

B. (i) inherent satisfaction of posting; (ii) decrease the number of user posts

C. (i) posting motivated by the perceptions of others; (ii) increase the number of user posts

D. (i) posting motivated by the perceptions of others; (ii) decrease the number of user posts

Question 5.
Which of the following statements best describes the primary difference between Difference-in-Differences (DiD) and CUPED (Controlled-experiment Using Pre-Experiment Data) research designs?

A. DiD uses differences in pre-treatment trends to control for time-varying confounders, whereas CUPED uses pre-treatment data to adjust post-treatment outcomes for increased precision.

B. CUPED relies on creating synthetic control groups to estimate treatment effects, whereas DiD compares changes in outcomes over time between treated and untreated groups.

C. DiD is used in observational studies without randomized assignment, whereas CUPED is used in both experimental and quasi-experimental designs to reduce variance.

D. CUPED controls for selection bias by matching on pre-treatment characteristics, whereas DiD controls for time-invariant differences between treated and control groups.

Question 6.
A company uses observational data paired with linear regression to analyze the effect of ad exposure on purchase behaviour over a one month time period. Their findings reveal that users who see the ads are more likely to purchase the product. What is the primary reason why this linear regression should not be interpreted causally?

A. The model does not account for the seasonal variations in sales data.

B. The sample size is too small to draw meaningful conclusions.

C. Selection bias, because users who see the ads may differ systematically from those who do not.

D. The regression assumes a linear relationship between ad exposure and purchase behaviour, which might not be accurate.

Question 7.
What is the primary purpose of word stemming and word lemmatization when analyzing text data?

A. To remove stop words from text data to reduce dimensionality.

B. To convert words to their root or base form for standardizing text data.

C. To identify named entities within the text, such as names of people or places.

D. To detect the sentiment of a given text by analyzing the intensity of words.

Question 8.
In quantitative marketing research, what does the process of **identification** refer to?

A. The technique of selecting the appropriate model specification for analysis.

B. The method of gathering a representative sample to ensure generalizability.

C. The approach of differentiating between endogenous and exogenous variables.

D. The process of figuring out what part of the variation in your data answers the research question.

Question 9.
In a study assessing the impact of social media engagement (S) on website traffic (T), a researcher runs a regression model:

$$T_i = \beta_0 + \beta_1 S_i + \varepsilon_i$$

However, the researcher overlooks an important variable: content quality (Q) of social media posts. If content quality (Q) is omitted from the model, what bias is likely to affect the estimated effect of social media engagement (S) on website traffic (T)?

A. Overestimation of β_1 due to omitted variable bias as content quality (Q) is positively correlated with both social media engagement (S) and website traffic (T).

B. Underestimation of β_1 due to omitted variable bias, as content quality (Q) is negatively correlated with social media engagement (S) but positively correlated with website traffic (T).

C. Attenuation bias in β_1 as inaccuracies in measuring social media engagement (S) could bias the estimated coefficient.

D. Selection bias, as the researcher might not have included a representative sample of social media posts in the analysis.

Question 10.
Complete the sentence. The adoption of management responses to online reviews leads to:

A. No change in the length of subsequent reviews

B. An increase in the length of all subsequent reviews

C. An increase in the length of subsequent positive reviews

D. An increase in the length of subsequent negative reviews

Part B: Short Answer Questions (70 points)

Answer ***all*** questions in this section. Point allocations per question are clearly indicated by each question and/or sub-question.

Question 1 (14 points). Text Analytics.

When discussing this class with a friend, they are interested in learning more about text analytics. They ask you the following question:

“Can you explain to me the differences and similarities between topic modelling and sentiment analysis?”

Write a response to your friend, intuitively explaining the fundamental similarities and distinctions between sentiment analysis and topic modelling. Provide clear explanations supported by relevant examples. (max. 8 sentences)

Question 2 (14 points). Influencer Disclosure Regulation

(a) [4 points] What is the goal of influencer disclosure regulation? Why is it important? (max 3 sentences)

(b) [6 points] Explain how you could use the introduction of influencer disclosure regulation in Germany but not in Spain to estimate the effect of the regulation on the share of disclosed posts and the impact on consumer engagement (max. 6 sentences)

(c) [4 points] What is the direction of effect on the outcome variables mentioned above found by Ershov and Mitchell? What are the implications for managers and/or policymakers? (max. 4 sentences)

Question 3 (14 points). Donors Sharing About Their Charitable Contributions.

(a) [4 points] Explain the tradeoff that donors face when deciding to share about their charitable donations. (max 3 sentences)

(b) [6 points] Explain the research design a team of researchers used to test a new strategy to encourage greater sharing. (max 5 sentences)

(c) [4 points] Summarise the main findings of the research and explain their managerial relevance. (max 3 sentences)

Question 4 (14 points). A/B Test design

You work for a platform that allows users to rate and share products with their friends. Your team wants to evaluate a new “personalized product feed” using an A/B test, where some users (treatment group) receive the new feed, and others (control group) do not. However, users are embedded in a social network — and often influence each other’s behavior.

(a) [4 points] Why might a standard A/B test fail to provide an unbiased estimate of the treatment effect in this setting? (max 3 sentences)

(b) [6 points] Describe an alternative experimental design that could help address the challenge identified in part (a). Justify your approach and explain how it improves the validity of the estimated effect. (max 6 sentences)

(c) [4 points] Suppose you run your proposed experiment and find that both treated users and some untreated users show increased engagement. What does this pattern suggest, and how would you interpret and report these findings to decision-makers? (max 3 sentences)

Question 5 (14 points). Topic Modelling.

(a) [5 points] Suppose you are given a dataset called `reviews_df` that contains one column, `review_text`, with raw hotel reviews. The dataset is **not yet tokenized**.

Explain the steps required to convert this dataset into a *Document-Term Matrix (DTM)* suitable for topic modeling. Include the key R functions or packages you would use at each step, and briefly justify why each step is necessary. (max 5 sentences)

(b) [4 points]. Assume that you have the data set of reviews about hotels that has already been cleaned and converted to a Document-Term-Matrix called `reviews_dtm`.

Write R code that estimates a structural topic model with 7 topics that would output the same output if run more than once.

(c) [5 points] Explain how a marketing analytics team that works for Hilton Hotels could use sentiment analysis together with topic modelling to propose new managerial strategies that improve the consumer experience. (max 5 sentences)

(END OF EXAM.)